

A Large Self-Annotated Corpus for Sarcasm

Mikhail Khodak and Nikunj Saunshi and Kiran Vodrahalli

Computer Science Department, Princeton University

35 Olden St., Princeton, New Jersey 08540

{mkhodak, nsaunshi, knv}@cs.princeton.edu

Abstract

We introduce the Self-Annotated Reddit Corpus (**SARC**)¹, a large corpus for sarcasm research and for training and evaluating systems for sarcasm detection. The corpus has 1.3 million sarcastic statements — 10 times more than any previous dataset — and many times more instances of non-sarcastic statements, allowing for learning in regimes of both balanced and unbalanced labels. Each statement is furthermore *self-annotated* — sarcasm is labeled by the author and not an independent annotator — and provided with user, topic, and conversation context. We evaluate the corpus for accuracy, compare it to previous related corpora, and provide baselines for the task of sarcasm detection.

1 Introduction

Sarcasm detection is an important component of many natural language processing (NLP) systems, with direct relevance to natural language understanding, dialogue systems, and text mining. However, detecting sarcasm is difficult because it occurs infrequently and is difficult for even human annotators to discern (Wallace et al., 2014). Despite these properties, existing datasets either have *balanced labels* — data with approximately the same number of examples for each label (González-Ibáñez et al., 2011; Bamman and Smith, 2015; Joshi et al., 2015; Amir et al., 2016; Oraby et al., 2016) — or use human annotators to label sarcastic statements (Riloff et al., 2013; Swanson et al., 2014; Wallace et al., 2015).

In this work, we make available the first corpus for sarcasm detection that has both unbalanced and

self-annotated labels and does not consist of low-quality text snippets from Twitter². With more than a million examples of sarcastic statements, each provided with author, topic, and context information, the dataset also exceeds all previous sarcasm corpora by an order of magnitude. This dataset is possible due to the comment structure of the social media site Reddit³ as well its frequently-used and standardized annotation for sarcasm.

Following a discussion of corpus construction and relevant statistics, in Section 4 we present results of a manual evaluation on a subsample of the data as well as a direct comparison with alternative sources. Then in Section 5 we examine simple methods of detecting sarcasm on both a balanced and unbalanced version of our dataset.

2 Related Work

Since our main contribution is a corpus and not a method for sarcasm detection, we point the reader to a recent survey by Joshi et al. (2016) that discusses many interesting efforts in this area. Note that many of the works the authors mention will be discussed by us in this section, with many papers using their own datasets; this illustrates the need for common baselines for evaluation.

Sarcasm datasets can largely be distinguished by the sources used to get sarcastic and non-sarcastic statements, the amount of human annotation, and whether the dataset is balanced or unbalanced. Reddit has been used before, notably by Wallace et al. (2015); while the authors allow unbalanced labeling, they do not exploit the possibility of using self-annotation and generate around 10,000 human-labeled sentences. Twitter is a frequent source due to the self-annotation provided by hashtags such as #sarcasm, #notsarcasm, and

¹<http://nlp.cs.princeton.edu/SARC/>

²<https://www.twitter.com>

³<https://www.reddit.com>

#irony (Reyes et al., 2013; Bamman and Smith, 2015; Joshi et al., 2015). As discussed in Section 4.2, its low quality language and other properties make Twitter a less attractive source for annotated comments. However, it is by far the largest raw source of data for this purpose and has led to some large unbalanced corpora in previous efforts (Riloff et al., 2013; Ptáček et al., 2014). A further source of comments is the Internet Argument Corpus (IAC) (Walker et al., 2012), a scraped corpus of Internet discussions that can be further annotated for sarcasm by humans or by machine learning; this is done by Lukin and Walker (2013) and Oraby et al. (2016), in both cases resulting in around 10,000 labeled statements.

3 Corpus Details

3.1 Reddit Structure and Annotation

Reddit is a social media site in which users communicate by commenting on *submissions*, which are titled posts consisting of embedded media, external links, and/or text, that are posted on topic-specific forums known as *subreddits*; examples of subreddits include *funny*, *pics*, and *science*. Users comment on submissions and on other comments, resulting in tree-like conversation structure such that each comment has a parent comment. We refer to *elements* as any nodes in the tree of a Reddit link (i.e., comments or submissions).

Users on Reddit have adopted a common method for sarcasm annotation consisting of adding the marker “/s” to the end of sarcastic statements; this originates from the HTML text delination `<sarcasm>...</sarcasm>`. As with Twitter hashtags, using these markers as indicators of sarcasm is noisy (Bamman and Smith, 2015), especially since many users do not make use of the marker, do not know about it, or only use it where sarcastic intent is not otherwise obvious. We discuss the extent of this noise in Section 4.1.

3.2 Constructing SARC

Reddit comments from December 2005 have been made available due to web-scraping⁴; we construct our dataset as a subset of comments from 2009-2016, comprising the vast majority of comments and excluding noisy data from earlier years. For each comment we provide a sarcasm label, author, the subreddit it appeared in, the comment



Figure 1: A Reddit submission and one of its comments. Note the conventional self-annotation “/s” indicating sarcasm.

score as voted on by users, the date of the comment, and the parent comment or submission.

To reduce noise, we use several filters to remove noisy and uninformative comments. Many of these are standard preprocessing steps such as excluding URLs and limiting characters to be ASCII. To handle Reddit data, we also perform the following two filtering steps:

- Only use comments from an author starting from the first month in which that author used the standard sarcasm annotation. This ensures that the author knows the annotation and makes unlabeled sarcasm less likely.
- Exclude comments that are descendants of sarcastic comments in the conversation tree, as annotation in such cases is extremely noisy, with authors agreeing or disagreeing with the previously expressed sarcasm with their own sarcasm but often not marking.

We collect a very large corpus, **SARC-raw**, with around 500-600 million total comments, of which 1.3 million are sarcastic. To obtain a smaller, clear corpus with better-quality comments, we take a subset consisting of single sentences having between 2 and 50 tokens. This yields a smaller corpus of around 200 million comments and 600 thousand sarcastic comments, **SARC-main**. Finally, we take a subset of the latter corpus corresponding to comments in the subreddit *politics*, which is a very large and sarcastic subreddit; we call this last segment **SARC-pol**.

4 Corpus Evaluation

There are three major metrics of interest for evaluating our corpora: (1) size, (2) the proportion of sarcastic to non-sarcastic comments, and (3) the rate of false positives and false negatives. Of interest is also the quality of the text in the corpus and

⁴<http://files.pushshift.io/reddit>

Corpus	Dataset	Sarc.	Total
IAC	Joshi et al. ‘15	751	1502
	Oraby et al. ‘16	4.7K	9.4K
Twitter	Joshi et al. ‘16	4.2K	5.2K
	Bamman & Smith ‘15	9.7K	19.5K
	Reyes et al. ‘13	10K	40K
	Riloff et al. ‘13	35K	175K
	Ptáček et al. ‘13	130K	780K
Reddit	Wallace et al. ‘15	753	14124
	SARC-pol	30K	4600K
	SARC-main	.63M	210M
	SARC-raw	1.4M	564M

Table 1: SARC compared with other corpora. Our dataset includes a million sarcastic comments.

its applicability to other NLP tasks. Thus in this section we consider evaluate error in the SARC-main subset and provide comparison with other corpora used to construct sarcasm datasets.

4.1 Manual Evaluation

To investigate the noisiness of using Reddit as a source of self-annotated sarcasm we estimate the proportion of false positives and false negatives induced by our filtering. This is done by having three human evaluators manually check a random subset of 500 comments from SARC-main tagged as sarcastic and 500 tagged as non-sarcastic, with full access to the comment’s context. A comment was labeled a false positive if a majority determined that the “/s” tag was not an annotation but part of the sentence and a false negative if a majority determined that the comment author was clearly being sarcastic. After evaluation, the false positive rate was determined to be 2.0% and the false negative rate 3.0%. Although the false positive rate is reasonable, the false negative rate is significant compared to the sarcasm proportion, indicating large variation in the working definition of sarcasm and the need for methods that can handle noisy data in the unbalanced setting.

4.2 Comparison with other Sources

As noted before, Twitter has been the most common source for sarcasm in previous corpora; this is likely due to the explicit annotation provided by its hashtags. However, using Reddit as a source of sarcastic comments holds many research advantages. Unlike Reddit comments, which are not

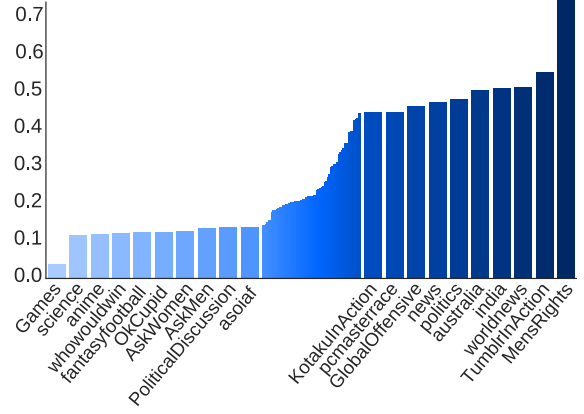


Figure 2: Sarcasm percentage for subreddits with more than a million comments in SARC-raw. Well-moderated and special-interest forums such as science and asoiat (referring to fantasy series *A Song of Ice and Fire*) have less sarcasm than controversial and less-moderated subreddits.

constrained by length and contain fewer hashtags, tweets are not written in true English. Hashtagged tokens are also frequently used as a part of the statement itself (e.g. “that was #sarcasm”), blurring the line between text and annotation; on Reddit “/s” is generally only used as something other than annotation when its use as an annotation is being referred to (e.g. “you forgot the /s”).

Furthermore, from a subsample of Twitter and Reddit data from July 2014 we determined that a vastly smaller percentage (.002% vs. .927%) of Twitter authors make use of sarcasm annotation (#sarcasm, #sarcastic, or #sarcastictweet). We hypothesize that Reddit users require sarcastic annotation more frequently and in a more standardized form because they are largely anonymous and so cannot rely on a shared context to communicate sarcasm. Finally, Reddit also benefits from having subreddits, which enable featurization and data exploration based on an explicit topic assignment.

The Internet Argument Corpus (IAC) has also been used as a source of sarcastic comments (Walker et al., 2012). The corpus developers found 12% of examples in the IAC to be sarcastic, which is a much nicer class proportion for sarcasm detection than ours. As the Reddit data consists of arbitrary conversations, not just arguments, it is not surprising that our sarcasm percentage is much smaller, even when accounting for false negatives; this property also makes our dataset more realistic. Unlike Reddit and Twitter, the IAC also requires manual annotation of sarcasm.

Regime	Method	SARC-main			SARC-pol		
		F_1	Precision	Recall	F_1	Precision	Recall
Unbalanced	Bag-of-NGrams	5.8	4.9	6.9	10.0	10.3	9.6
Balanced	Bag-of-NGrams	73.1	75.0	71.3	73.6	75.3	72.0
	Average Embedding	64.6	65.6	63.6	67.7	69.1	66.4
	SNIF Embedding	65.1	66.9	63.4	68.7	70.9	66.6
	Human*	71.9	84.0	62.8	71.7	80.5	64.6

* Majority decision of three independent fluent English speakers.

Table 2: Baseline Performance for Sarcasm Detection

5 Baselines for Sarcasm Detection

A direct application of our corpus is for training and evaluating sarcasm detection systems; we thus provide baselines for the task of classifying a given statement as sarcastic or not-sarcastic. We consider non-contextual representations of the comments as feature vectors to be classified by a regularized linear support vector machine (SVM) trained by stochastic gradient descent (SGD). In the unbalanced regime we use validation to tune the class-weight, assigning more weight to the sarcasm class to force its consideration in the loss.

5.1 Bag-of-NGrams

The Bag-of-NGrams representation consists of using a document’s n -gram counts as features in a vector. For the SARC-main subset we use all unigrams, bigrams, and trigrams that appear at least 5 times in the sarcastic training comments. For the SARC-raw subset we use all unigrams, bigrams, and trigrams that appear at least 5 times in all training comments. The feature vectors are normalized before classification.

5.2 Word Embeddings

Given a document, taking the elementwise average of word embeddings of its words provides a simple low-dimensional document representation. For word vectors we use normalized 300-dimensional GloVe representations trained on the Common Crawl corpus (Pennington et al., 2014). Since we are establishing baselines and SGD yields inconsistent results on word embedding representations in the unbalanced regime, we do report not any word embedding results for that task.

Smooth Inverse Frequency (SIF) embedding, in which a document is represented as a weighted average of the embeddings of its words, has been shown to be an effective baseline representation

speculation is so much more fun than all those pesky facts , amirite ?

yeah , totally nothing wrong with bullying people with the threat of meritless litigation

Figure 3: Two example sarcastic sentences from SARC-main. Darker shading corresponds to non-stopwords with a higher SNIF weight.

method (Arora et al., 2017). Given a word w ’s frequency f_w , SIF-weights assign a weight $\frac{a}{a+f_w}$, where a is a hyperparameter often set to $a = 10^{-4}$. As in TF-IDF, this assigns a low weight to high frequency words; however, it performs surprisingly poorly compared to average embedding.

Hypothesizing that, unlike in regular text classification, a document’s sarcasm depends on words with fairly high frequencies such as *sure*, *totally*, and *wow*, we instead use Smooth *Negative Inverse Frequency* (SNIF) weights, assigning $1 - \frac{a}{a+f_w}$ to each word before also representing the document as a weighted average of word embeddings. We find that such a weighting does in fact improve performance over vector averaging.

6 Conclusion

We introduce a large sarcasm dataset based on self-annotated Reddit comments. The datasets are provided in three versions, allowing for diverse applications: **SARC-raw**, **SARC-main**, and **SARC-pol**. **SARC-raw** has over 1 million sarcastic sentences, larger than any existing dataset. We evaluate the baseline performance of simple machine learning methods and compare them with human performance, with SVM over Bag-of-NGram representations outperforming humans on the task of sarcasm detection. We hope that future users of this dataset will improve upon these benchmarks.

7 Acknowledgements

We would like to acknowledge Angel Chang and Christiane Fellbaum for helpful discussion as well as Amy Hua and Elizabeth Yang for help with evaluation.

References

- Silvio Amir, Byron C. Wallace, Hao Lyu, Paula Carvalho, and Mário J. Silva. 2016. Modelling context with user embeddings for sarcasm detection in social media. <https://arxiv.org/pdf/1607.00976.pdf>.
- Sanjeev Arora, Yingyu Liang, and Tengyu Ma. 2017. A simple but tough-to-beat baseline for sentence embeddings. Conference Paper at ICLR 2017. <https://openreview.net/pdf?id=SyK00v5xx>.
- David Bamman and Noah A. Smith. 2015. Contextualized sarcasm detection on twitter. Association for the Advancement of Artificial Intelligence.
- Roberto González-Ibáñez, Smaranda Muresan, and Nina Wacholder. 2011. Identifying sarcasm in twitter: A closer look. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, pages 581–586.
- Aditya Joshi, Pushpak Bhattacharyya, and Mark J. Carman. 2016. Automatic sarcasm detection: A survey. <https://arxiv.org/pdf/1602.03426.pdf>.
- Aditya Joshi, Vinita Sharma, and Pushpak Bhattacharyya. 2015. Harnessing context incongruity for sarcasm detection. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, pages 757–762. <http://www.aclweb.org/anthology/P15-2124>.
- Stephanie Lukin and Marilyn Walker. 2013. Really? well, apparently bootstrapping improves the performance of sarcasm and nastiness classifiers for on-line dialogue. In *Proceedings of the Workshop on Language in Social Media*. Association for Computational Linguistics, pages 30–40.
- Shereen Oraby, Vrindavan Harrison, Lena Reed, Ernesto Hernandez, Ellen Riloff, and Marilyn Walker. 2016. Creating and characterizing a diverse corpus of sarcasm in dialogue. In *Proceedings of the SIGDIAL 2016 Conference*. Association for Computational Linguistics, pages 31–41.
- Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, page 15321543. <http://www.aclweb.org/anthology/D14-1162>.
- Tomáš Ptáček, Ivan Habernal, and Jun Hong. 2014. Sarcasm detection on czech and english twitter. In *25th International Conference on Computational Linguistics: Technical Papers*. pages 213–223. <http://www.aclweb.org/anthology/C14-1022>.
- Antonio Reyes, Paolo Rosso, and Tony Veale. 2013. A multidimensional approach for detecting irony in twitter. *Data Knowledge Engineering* 47(1).
- Ellen Riloff, Ashequl Qadir, Prafulla Surve, Lalin-dra De Silva, Nathan Gilbert, and Ruihong Huang. 2013. Sarcasm as contrast between a positive sentiment and negative situation. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, pages 704–741. <http://www.aclweb.org/anthology/D13-1066>.
- Reid Swanson, Stephanie Lukin, Luke Eisenberg, Thomas Chase Corcoran, and Marilyn A. Walker. 2014. Getting reliable annotations for sarcasm in online dialogues. In *Language Resources and Evaluation Conference*.
- Marilyn A. Walker, Pranav Anand, Jean E. Fox Tree, Rob Abbott, and Joseph King. 2012. A corpus for research on deliberation and debate. In *Language Resources and Evaluation Conference*.
- Byron C. Wallace, Do Kook Choe, and Eugene Charniak. 2015. Sparse, contextually informed models for irony detection: Exploiting user communities, entities, and sentiment. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, pages 1035–1044. <https://doi.org/10.18653/v1/P15-1100>.
- Byron C. Wallace, Do Kook Choe, Laura Kertz, and Eugene Charniak. 2014. Humans require context to infer ironic intent (so computers probably do, too). In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, pages 512–516.